Bayesian Blind Deconvolution with General Sparse Image Priors

S. Derin Babacan[1], Rafael Molina[2], Minh N. Do[1], and Aggelos K. Katsaggelos[3]

[1] University of Illinois at Urbana-Champaign
[2] Universidad de Granada
[3] Northwestern University

Abstract. We present a general method for blind image deconvolution using Bayesian inference with super-Gaussian sparse image priors. We consider a large family of priors suitable for modeling natural images, and develop the general procedure for estimating the unknown image and the blur. Our formulation includes a number of existing modeling and inference methods as special cases while providing additional flexibility in image modeling and algorithm design. We also present an analysis of the proposed inference compared to other methods and discuss its advantages. Theoretical and experimental results demonstrate that the proposed formulation is very effective, efficient, and flexible.

1 Introduction

Blind image deconvolution is the problem of restoring an image x from its blurred and noisy version y when the blur kernel k is unknown. Generally, the image y is modeled as the convolution of the unknown sharp image x with the blur kernel as

$$y = k \otimes x + n \tag{1}$$

where n is the noise. Since k, x and n are unknown, the problem is highly ill-posed and there are infinitely many solutions for x and k. Moreover, in most cases the blur kernel k is spatially-varying. To obtain meaningful solutions, the problem must be regularized with additional information about the image x, noise n, and kernel k. This regularization is generally embedded by assigning priors $p(x)$ and $p(k)$ which reflect our prior knowledge on the characteristics of x and k.

Blind image deconvolution is a widely investigated problem in signal/image processing and computer vision [1], and recently attracted much attention mostly geared towards removing camera shake [2–10]. Fergus *et. al.* [2] employed the variational Bayesian approach of Miskin and Mackay [11] with a mixture-of-Gaussians image prior for modeling natural image statistics. After the success of this approach, subsequent methods proposed new image and blur modeling schemes and highly efficient inference methods [5–8]. While early approaches

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 341–355, 2012.

assume a spatially-invariant blur, more recent approaches addressed the more general and challenging problem of removing spatially varying blurs [10, 12, 13].

The blind deconvolution problem contains two interacting components: modeling and inference. In modeling, a commonly used principle is that natural images have super-Gaussian statistics, and blur causes the statistics to become more similar to Gaussian by smoothing out sharp gradients. Hence, deconvolution should make the statistics less Gaussian, which in turn leads to the common use of super-Gaussian (or sparse) image priors. Unfortunately, direct use of these priors makes the second part of blind deconvolution, inference, challenging, in most cases limiting the options to maximum *a posteriori* (MAP) estimation. However, it has been shown in [8] that MAP is not suitable for blind deconvolution. The problem stems from the fact that while sharp images are well modeled with super-Gaussian priors, blurred images are also relatively well modeled with these priors. Hence, the prior alone is not sufficient to force the algorithm to choose a sharp image, and MAP generally leads to the trivial no-blur solution. Based on this observation, [8, 9] advocated Bayesian inference methods where the image is marginalized from the optimization while estimating the unknown blur. While [9] showed the advantages of such schemes using Gaussian and some sparse image priors, Bayesian inference for blind deconvolution with super-Gaussian priors remains a challenging obstacle.

In this paper, we present a general formulation for blind deconvolution using sparse image priors from both modeling and inference perspectives. For image modeling, we introduce a large class of sparse image priors suitable for representing sharp image characteristics. Most models used in the literature are included in our formulation as special cases while we propose a new and powerful alternative. Using this general prior formulation, we develop the estimation procedure for the unknown image and blur kernel using variational Bayesian inference. We analyze the proposed inference method in comparison with MAP, and show that MAP estimation is generally not suitable for inference for both blind and non-blind deconvolution. The proposed inference naturally addresses its shortcomings due to an implicit regularization mechanism with minimal additional algorithmic complexity. Finally, we demonstrate that this formulation of super-Gaussian priors and inference leads to methods for *constructing* image priors instead of using limited parametric forms, which provides additional flexibility in image modeling and algorithm design for blind deconvolution.

In the following, for notational simplicity, we treat images with N pixels as $N \times 1$ vectors and denote T_k as the matrix operator implementing $T_k\, x = k \otimes x$. $\mathrm{diag}\,(\cdot)$ creates a diagonal matrix from its argument.

2 A General Sparse Prior Model for Deconvolution

The deconvolution problem can be formulated in either filter or image space [8]. In the filter space approach, we create L pseudo-observations y_γ by applying high-pass filters $\{f_\gamma\}_{\gamma=1}^{L}$ (such as derivatives, wavelets, curvelets, etc.) to the blurred noisy image y as

$$y_\gamma = f_\gamma \otimes y = k \otimes f_\gamma \otimes x + f_\gamma \otimes n = k \otimes x_\gamma + n_\gamma \tag{2}$$

with $x_\gamma = f_\gamma \otimes x$. The image priors are placed on the filtered image coefficients $\{f_\gamma \otimes x\}_{\gamma=1}^{L}$, hence L image priors $\{p(x_\gamma)\}_{\gamma=1}^{L}$ are defined. In the image space approach, we define a single prior directly on the unknown image, albeit by using its filtered coefficients. In this article, we follow the filter space approach as it is considerably simpler, but the main principles apply to the image space approach as well. Finally, since we estimate the filtered images x_γ, a crucial question is how to infer x from x_γ's, for which we will provide an effective method later.

It is well known that when high-pass filters are applied to natural images, the resulting coefficients are *sparse*; i.e., most of the coefficients are zero or very small while only a small number of coefficients are large (e.g., at the edges). This behavior is exploited in all advanced blind deconvolution methods using sparse image priors. An image prior is considered to be sparse when it is super-Gaussian [14], i.e., compared to the Gaussian distribution, it has heavier tails, it is more peaked, and has a positive excess kurtosis. These distributions are referred to as sparse since most of the distribution mass is located around zero (hence strongly favoring zero values), but the probability of occurrence of large signal values is higher compared to the Gaussian distribution.

In this article, we consider the following general form of super-Gaussian image priors on x_γ.

$$p(x_\gamma) = \prod_{i=1}^{N} \mathrm{p}\left(x_\gamma(i)\right) = Z \exp\left(-\sum_i \rho\left(x_\gamma(i)\right)\right) \tag{3}$$

where Z is the normalization constant, $\rho(\cdot)$ is a penalty function (symmetric around 0), and $x_\gamma(i)$ denotes the filter output at pixel i. Sparsity is achieved when the function ρ leads to the suppression of most coefficients $x_\gamma(i)$ while preserving a small number of important features. Some examples are shown in Fig. 1. While some of these priors are commonly used in the deconvolution literature (like Gaussian and $\|\cdot\|_p^p$), we also introduce new ones (log and exp) not considered before. $\log|s|$ enforces sparsity very strongly due to its infinite peak at the origin and heavy tails, which proves very useful in kernel estimation.

In general, the prior (3) cannot be directly used in Bayesian inference. Next we describe variational representations to convert it to forms suitable for inference.

2.1 Variational Representations of Sparse Image Priors

Formally, for $p(x_\gamma)$ to be super-Gaussian, the function $\rho(\sqrt{s})$ has to be increasing and concave for $s \in (0, \infty)$ [14]. This condition is equivalent to $\rho'(s)/s$ being decreasing on $(0, \infty)$, that is, for $s_1 \geq s_2 \geq 0$, $\rho'(s_1)/s_1 \leq \rho'(s_2)/s_2$. If this condition is satisfied, then ρ can be represented as (using [15, Ch. 12])

$$\rho\left(x_\gamma(i)\right) = \inf_{\xi_\gamma(i)>0} \frac{1}{2}\,\xi_\gamma(i)\,x_\gamma^2(i) - \rho^*\left(\frac{1}{2}\xi_\gamma(i)\right) \tag{4}$$

$$\Rightarrow \rho\left(x_\gamma(i)\right) \leq \frac{1}{2}\,\xi_\gamma(i)\,x_\gamma^2(i) - \rho^*\left(\frac{1}{2}\xi_\gamma(i)\right) \tag{5}$$

$\rho(s)$	$\rho'(s)/\lvert s\rvert$
$\frac{1}{2}\lvert s\rvert^2$	1
$\frac{1}{p}\lvert s\rvert^p$	$\lvert s\rvert^{p-2}$
$\log \lvert s\rvert$	$\lvert s\rvert^{-2}$
$-\sigma_r \exp\left(-\frac{\lvert s\rvert^2}{2\sigma_r}\right)$	$\exp\left(-\frac{\lvert s\rvert^2}{2\sigma_r}\right)$

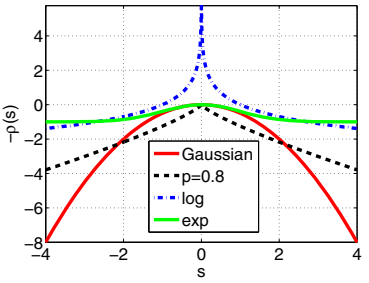

Fig. 1. Some choices of the penalty function ρ (left), and the corresponding penalties -ρ (right). The exp curve is vertically shifted and $\log|s|$ is bounded for better visualization.

where inf denotes the infimum, $\rho^*\left(\xi_\gamma(i)/2\right)$ is the concave conjugate of $\rho(\sqrt{x_\gamma(i)})$ and $\xi_\gamma = \{\xi_\gamma(i)\}_{i=1}^N$ are variational parameters. These parameters have an intuitive meaning and extreme importance in the deconvolution performance, as will be shown later. The relationship dual to (4) is given by [15]

$$\rho^*\left(\frac{1}{2}\xi_\gamma(i)\right) = \inf_{x_\gamma(i)} \frac{1}{2}\,\xi_\gamma(i)\,x_\gamma^2(i) - \rho\left(x_\gamma(i)\right) . \tag{6}$$

The quadratic bound for ρ in (5) allows us to bound the prior with a Gaussian form. Specifically, we can rewrite (3) as

$$p(x_\gamma) \geq Z \exp\left(-\frac{1}{2}\sum_i \xi_\gamma(i)\,x_\gamma^2(i)\right) \exp\left(\sum_i \rho^*\left(\frac{1}{2}\xi_\gamma(i)\right)\right) . \tag{7}$$

Equality in (7) is obtained at the optimal values of $\xi_\gamma(i)$, which are computed from the dual representation (6) by taking the derivative with respect to $x_\gamma(i)$ and setting it to zero, which gives $\xi_\gamma(i) = \rho'(x_\gamma(i))/|x_\gamma(i)|$.

This representation based on bounding has been widely used in computer vision (see, e.g., [16]). Another representation is given by

$$p(x_\gamma) = \int p(x_\gamma|\xi_\gamma)\,p(\xi_\gamma)\,d\xi_\gamma \tag{8}$$

where $p(x_\gamma|\xi_\gamma)$ is a Gaussian distribution with variance ξ_γ^{-1}. This is the well known scale mixture of Gaussians (SMG) [17], which defines x_γ using $x_\gamma(i) = \xi_\gamma(i)^{-\frac{1}{2}}\,z$ with z a standard Gaussian variable with zero mean and unit variance. This representation is more strict than (7) in the sense that a (slightly) more limited class of priors can be represented using (8)[1]. Finding $p(\xi_\gamma)$ is also in general much harder than ρ^*, but neither of them are needed for our purposes (see Sec. 3.1). All priors in Fig. 1 can be represented both ways. We will use

[1] The scale mixture of Gaussian representation requires complete monotonicity of $p(\sqrt{s})$ [17]. A function $f(s)$ is completely monotonic if its derivatives satisfy $(-1)^n f^{(n)}(s) \geq 0$ for all $n = 0, 1, 2, ...$

the SMG representation for deriving the VB inference, and (7) for deriving the related cost function formulation in Sec. 3.2. The explicit link between these in inference will be shown in a subsequent article.

Using these representations, we are able to transform the super-Gaussian priors to Gaussian forms, rendering the optimization much easier compared to the original forms. This is achieved by expanding the optimization with respect to (w.r.t.) x_γ to joint optimization w.r.t. x_γ and ξ_γ. However, it should be emphasized that a super-Gaussian prior is enforced only when these parameters are jointly estimated, which shows the tight coupling between the modeling and the inference procedures.

Most priors used in blind deconvolution are obtained as special cases of the formulation in (4), including priors based on l_p-norms (hyper-Laplacian) in [3,7, 18], and the piecewise linear/quadratic in [5]. The Gaussian prior used in [6,13] is also a special (limiting) case where ξ_γ is set to a constant. We will show later that other priors such as mixture-of-Gaussians can also be obtained from (4).

Having defined a general class of sparse image priors, we define the distribution of the observed images y_γ by assuming white Gaussian noise for n_γ as

$$p(y_\gamma|x_\gamma,k) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\|y_\gamma - k\otimes x_\gamma\|_2^2\right) \tag{9}$$

where σ^2 is the noise variance. We do not use a specific prior for the blur kernel k, i.e., $p(k) = \text{const}$, but we will constrain it as discussed later. Finally, notice that we have defined L distributions for the latent images x_γ and the observations y_γ. For clarity, in the following we denote the joint distributions using p_T, e.g., $p_T(y_\gamma|x_\gamma,k) = \prod_{\gamma=1}^{L} p(y_\gamma|x_\gamma,k)$.

3 Inference

Virtually all inference schemes are based on the posterior distribution, which in our modeling takes the form (using the SMG representation (8))

$$p_T(x_\gamma,\xi_\gamma,k|y_\gamma) = \frac{p_T(y_\gamma,x_\gamma,\xi_\gamma,k)}{p_T(y_\gamma)} = \frac{p_T(y_\gamma|x_\gamma,k)\,p_T(x_\gamma|\xi_\gamma)\,p_T(\xi_\gamma)\,p(k)}{p_T(y_\gamma)}\,. \tag{10}$$

However, the exact posterior cannot be calculated since the required integration $p(y_\gamma) = \int p(y_\gamma,x_\gamma,\xi_\gamma,k)\,dx_\gamma\,d\xi_\gamma\,dk$ is analytically intractable, as is the case with almost all deconvolution methods.

Since the posterior distribution cannot be obtained, a crucial question is to decide how inference should be carried out. Perhaps the most commonly used inference scheme is the MAP approach, where the unknowns are estimated by the maximum (the mode) of the posterior. However, as we will show later, this approach is generally not suitable both for blind and non-blind deconvolution. In this article, we base our estimation on a general inference procedure, variational Bayes (VB) [19,20], where the intractable posterior is approximated with a tractable distribution. We chose VB since it overcomes the problems associated

with the MAP method while achieving almost the same computational and implementation complexity. Moreover, it includes several other inference strategies (such as MAP and expectation-maximization) as special cases.

3.1 Estimation Using Variational Bayes (VB)

In the variational Bayes approach, one approximates the posterior $p_T(x_\gamma, \xi_\gamma, k|y_\gamma)$ with another distribution $q_T(x_\gamma, \xi_\gamma, k)$ by minimizing the Kullback-Leibler (KL) divergence between them. This minimization becomes tractable if a suitable factorization for $q(x_\gamma, k, \xi_\gamma)$ is chosen. A convenient selection is the fully factorized distribution $q(x_\gamma, k, \xi_\gamma) = q(x_\gamma)\, q(k)\, q(\xi_\gamma)$, corresponding to the mean field approximation [19, 20]. In this case, the optimal q distribution for each variable is found by minimizing the KL divergence while holding the other q distributions fixed. It can be shown [19] that the optimal q distributions are obtained by taking the expectation of the joint distribution with respect to all unknowns except the one of interest, which in our modeling leads to

$$\begin{aligned}
\log q(x_\gamma) &= \mathrm{E}\left[\log p(x_\gamma, k, \xi_\gamma, y_\gamma)\right]_{\xi_\gamma, k} + \text{const} \\
&= \mathrm{E}\left[\log p(y_\gamma|x_\gamma, k)\right]_k + \mathrm{E}\left[\log p(x_\gamma|\xi_\gamma)\right]_{\xi_\gamma} + \text{const} \qquad (11) \\
\log q(k) &= \mathrm{E}\left[\log p_T(x_\gamma, k, \xi_\gamma, y_\gamma)\right]_{x_\gamma, \xi_\gamma} = \mathrm{E}\left[\log \mathrm{p}_T(y_\gamma|x_\gamma, k)\right]_{x_\gamma} + \text{const} \qquad (12) \\
\log q(\xi_\gamma) &= \mathrm{E}\left[\log p(x_\gamma, k, \xi_\gamma, y_\gamma)\right]_{x_\gamma, k} \\
&= \log p(\xi_\gamma) + \mathrm{E}\left[\log p(x_\gamma|\xi_\gamma)\right]_{x_\gamma} + \text{const}\,. \qquad (13)
\end{aligned}$$

The expectations are taken with respect to the corresponding q distributions. The estimates of the image, blur and the variational parameters are then taken as the mean values of these distributions, which we derive next.

Estimation of Image and Blur. For the blur, we can directly use (12) and estimate it using the mean of $q(k)$. However, this way we are not imposing any constraints on the blur kernel, since no informative prior is assigned on k. To include typical constraints $k \geq 0$ and $\sum_i k(i) = 1$, we can assign a Dirichlet prior on k, but this leads to a complicated inference procedure. We take another approach and treat k as a deterministic parameter, and simply minimize the KL divergence w.r.t. k subject to these constraints. In this approach, $q(k)$ becomes a delta distribution at the estimate $\hat{k}$, which is obtained by solving

$$\hat{k} = \arg\min_k \sum_\gamma \mathrm{E}\left[\|y_\gamma - k \otimes x_\gamma\|_2^2\right]_{x_\gamma} = \arg\min_k\; k^T C_k^{-1} k - 2\, k^T\, b_k \qquad (14)$$

subject to $k \geq 0$ and $\sum_i k(i) = 1$. For a kernel size of $M \times M$, b_k is $M^2 \times 1$ and matrix C_k^{-1} is $M^2 \times M^2$ with

$$C_k^{-1}(m,n) = \frac{1}{\sigma^2} \sum_{\gamma} \sum_{j=1}^{N} \mathrm{E}\left[x_\gamma(m+j)\right] \mathrm{E}\left[x_\gamma(n+j)\right] + C_{x_\gamma}(m+j, n+j) \quad (15)$$

$$b_k(m) = \frac{1}{\sigma^2} \sum_{\gamma} \sum_{j=1}^{N} \mathrm{E}\left[x_\gamma(m+j)\right]\, y_\gamma(j)\,. \quad (16)$$

The estimation of blur in (14) is thus a simple quadratic program and can be solved very efficiently.

For the image, we use the distribution $q(x_\gamma)$ in (11) which has a multivariate Gaussian form given by

$$-2\log q(x_\gamma) = x_\gamma^T C_{x_\gamma}^{-1} x_\gamma - 2\, b_{x_\gamma}^T\, x_\gamma + \text{const} \quad (17)$$

with

$$C_{x_\gamma}^{-1} = \frac{1}{\sigma^2} T_k^T T_k + \mathrm{diag}\left(\mathrm{E}\left[\xi_\gamma\right]\right), \qquad b_{x_\gamma} = \frac{1}{\sigma^2} T_k^T y \quad (18)$$

where C_{x_γ} is the covariance matrix of x_γ. The mean $\mathrm{E}\left[x_\gamma\right]$ of this distribution is used as the estimate for x_γ, which is obtained by solving the linear system $C_{x_\gamma}^{-1} \mathrm{E}\left[x_\gamma\right] = b_{x_\gamma}$ using the conjugate gradient method. Hence C_{x_γ} need not be formed explicitly, but it is required in (15). Since this computation is extremely expensive (it requires an $N \times N$ matrix inversion for an N-pixel image), we approximate it in (15) by a diagonal matrix by inverting only the diagonals of $C_{x_\gamma}^{-1}$, similarly to [9, 21] (see [21] for other approximations).

Estimation of Variational Parameters $\boldsymbol{\xi_\gamma}$. Finally, we need to calculate the distribution $q(\xi_\gamma)$ using (13), but this requires the calculation of $p(\xi_\gamma)$, which is generally hard. However, the full distribution $q(\xi_\gamma)$ is not needed to estimate x_γ and k; only the mean $\mathrm{E}\left[\xi_\gamma\right]$ is required to calculate C_{x_γ}. The mean is given by

$$\mathrm{E}\left[\xi_\gamma\right] = \int \xi_\gamma\, q(\xi_\gamma)\, d\xi_\gamma = \int \xi_\gamma\, p\left(\xi_\gamma | x_\gamma = \nu_\gamma\right) d\xi_\gamma \quad (19)$$

where $\nu_\gamma(i) = \sqrt{\left(\mathrm{E}\left[x_\gamma\right](i)\right)^2 + C_{x_\gamma}(i,i)}$. To calculate the integral, we examine

$$p'(x_\gamma) = \frac{\partial}{\partial x_\gamma} \left[\int p(x_\gamma|\xi_\gamma)\, p(\xi_\gamma)\, d\xi_\gamma\right] \quad (20)$$

$$\rho'(x_\gamma)\, p(x_\gamma) = \int \xi_\gamma\, x_\gamma\, p(x_\gamma)\, p(\xi_\gamma|x_\gamma) d\xi_\gamma \quad (21)$$

$$\Rightarrow \frac{\rho'(x_\gamma)}{x_\gamma} = \int \xi_\gamma\, p(\xi_\gamma|x_\gamma)\, d\xi_\gamma\,, \quad (22)$$

where the derivative and division are to be understood element-wise. We now see that (19) is equivalent to (22) when $x_\gamma = \nu_\gamma$, such that we have

$$\mathrm{E}\left[\xi_\gamma\right] = \frac{\rho'(\nu_\gamma)}{\nu_\gamma}\,. \quad (23)$$

Algorithm 1. Blind Deconvolution using General Sparse Image Priors

Inputs: Noisy and blurred image y, choice for ρ or κ function, filters f_γ.
Initialization: Set $x_\gamma = y_\gamma$, $C_{x_\gamma} = 0$, $\nu_\gamma = x_\gamma$.
while not converged **do**
1. Compute variational parameters ξ_γ using $\mathrm{E}\left[\xi_\gamma(i)\right] = \rho'(\nu_\gamma(i))/\nu_\gamma(i)$.
2. Estimate filtered images x_γ by solving $C_{x_\gamma}^{-1}\mathrm{E}\left[x_\gamma\right] = b_{x_\gamma}$ with C_{x_γ} and b_{x_γ} in (18).
3. Approximate $C_{x_\gamma}(i,i)$ with $1/C_{x_\gamma}^{-1}(i,i)$.
4. Set $\nu_\gamma(i) = \sqrt{\left(\mathrm{E}\left[x_\gamma\right](i)\right)^2 + C_{x_\gamma}(i,i)}$.
5. Estimate the blur kernel k using (14) .
end while
6. Compute the final image estimate $\hat{x}$ by solving

$$\left(T_k^T T_k + \sigma^2 \sum_\gamma T_{f_\gamma}^T \mathrm{diag}\left(\mathrm{E}\left[\xi_\gamma\right]\right) T_{f_\gamma}\right) \hat{x} = T_k^T\, y\,. \tag{24}$$

Summary. The proposed method is outlined in Algorithm 1. In summary, the method alternates between the estimates of x_γ, $C_{x_\gamma}(i,i)$, ξ_γ and k. Since the method estimates the filtered images x_γ, we need to construct the image x from x_γ. However, this is not trivial and requires careful integration of all x_γ. Instead, we propose to use the estimate shown in (24), which estimates x from the already estimated k and ξ_γ values, and still enforces sparsity in the filter domain through the use of ξ_γ. In addition, it requires only one more application of the conjugate gradient algorithm so it is computationally efficient.

3.2 A Cost-Function View of VB

Most blind deconvolution methods in the literature define a cost function which is minimized w.r.t. the unknown image and kernel. On the other hand, the VB approach employed here relies on a completely different strategy; it approximates the whole posterior distribution instead of point estimation. Here, we provide a cost function minimization formulation for the VB approach, which is helpful in explaining some of its properties and shows its advantages over MAP.

In the VB approach, the unknowns are estimated as the means of the approximating q distributions. Since the distribution of the filtered image $q(x_\gamma)$ in (17) is Gaussian, its mean coincides with its mode, and thus the VB estimate of x_γ in (18) can be expressed as the solution of

$$\mathrm{E}\left[x_\gamma\right] = \arg\min_{x_\gamma} \frac{1}{\sigma^2}\|y_\gamma - k \otimes x_\gamma\|_2^2 + x_\gamma^T \mathrm{diag}\left(\mathrm{E}\left[\xi_\gamma\right]\right) x_\gamma\,. \tag{25}$$

Combining this with the optimization problem for k from (14) and for ξ_γ from (23) (see the supplement), we obtain[2]

[2] Here, with an abuse of notation, C_{x_γ} is to be treated as constant during each alternating minimization, but then is updated with the new estimates.

$$\mathrm{E}\left[x_\gamma\right],\hat{k},\mathrm{E}\left[\xi_\gamma\right] = \underset{x_\gamma,k,\xi_\gamma}{\arg\min} \sum_\gamma \left[\frac{1}{\sigma^2}\|y_\gamma - k \otimes x_\gamma\|_2^2 + x_\gamma^T \mathrm{diag}\left(\xi_\gamma\right) x_\gamma \right.$$
$$\left. - 2\sum_i \rho^*(\xi_\gamma(i)/2) + \frac{1}{\sigma^2}\mathrm{trace}(T_k^T T_k C_{x_\gamma}) + \mathrm{trace}\left(\mathrm{diag}\left(\xi_\gamma\right) C_{x_\gamma}\right) \right]. \quad (26)$$

In comparison, the MAP estimation is formulated as

$$x_\gamma^{\mathrm{MAP}}, k^{\mathrm{MAP}}, \xi_\gamma^{\mathrm{MAP}} = \underset{x_\gamma,k,\xi_\gamma}{\arg\max}\ \log p_T(x_\gamma, k, \xi_\gamma | y_\gamma)$$
$$= \underset{x_\gamma,k,\xi_\gamma}{\arg\max}\ \log p_T(y_\gamma | x_\gamma, k)\ + \log p_T(x_\gamma|\xi_\gamma)\ + \log p_T(\xi_\gamma)$$
$$= \underset{x_\gamma,k,\xi_\gamma}{\arg\min} \sum_\gamma \left[\frac{1}{\sigma^2}\|y_\gamma - k \otimes x_\gamma\|_2^2 + \sum_i \xi_\gamma(i)\, x_\gamma^2(i) - 2\sum_i \rho^*(\xi_\gamma(i)/2) \right]. \quad (27)$$

Minimizing this objective in an alternating fashion yields

$$x_\gamma^{\mathrm{MAP}} = \left(\frac{1}{\sigma^2} T_k^T T_k + \mathrm{diag}\left(\xi_\gamma\right) \right)^{-1} \frac{1}{\sigma^2} T_k^T\, y \quad (28)$$

$$k^{\mathrm{MAP}} = \underset{k}{\arg\min}\ k^T \left(\frac{1}{\sigma^2} \sum_\gamma T_{x_\gamma}^T\, T_{x_\gamma} \right) k - \frac{1}{\sigma^2} T_{x_\gamma}^T\, y \quad (29)$$

$$\xi_\gamma(i)^{\mathrm{MAP}} = \rho'(x_\gamma(i)) \,/\, |x_\gamma(i)|\,. \quad (30)$$

Notice that both objective functions and all estimates in the VB and MAP approaches have exactly the same form, except for the inclusion of the term $C_{x_\gamma}(i,i)$ in VB, which is very easy to compute using the diagonal approximation. This difference is of high importance in estimation performance both in theory and practice, as shown next.

Examining the objective (27) shows that MAP is not suitable for joint estimation of x_γ, k and ξ_γ as the global minimum is obtained at the trivial solution $x_\gamma = 0$. This occurs because of two problems: First, the minimization w.r.t. x_γ and k with fixed ξ_γ results in the solution $x_\gamma = 0$, which can be shown similarly to [8] as follows: Consider a pair x_γ, k and define $x_\gamma^c = c\, x_\gamma$, $k^c = k\,/c$ with c a scalar. As $c \to 0$, the term $\|y_\gamma - k \otimes x_\gamma\|_2^2$ remains constant but the prior term $\sum_i \xi_\gamma(i) x_\gamma^2(i) c^2$ always decreases. On the contrary, the VB objective does not monotonically decrease as $c \to 0$, as C_{x_γ} is positive definite and $\mathrm{trace}(T_k^T T_k C_{x_\gamma})$ increases as k^c increases. Thus, the VB approach implicitly enforces a constraint on the kernel k even when no prior is defined on k.

Levin *et. al.* [8] showed that when k is constrained to $\sum_i k(i) = 1$, a no blur solution (a delta kernel) is preferred in the MAP approach. In the VB approach this effect is also stabilized: the term $\mathrm{trace}(T_k^T T_k C_{x_\gamma})$ forces $\sum_i k^2(i)$ to be small, which is equivalent to favoring kernel estimates with larger support (since $\sum_i k(i) = 1$). Hence, while the first two terms in (26) are pulling the estimates to a no blur solution, the second last term is balancing their effect.

Second, the MAP approach is problematic even with fixed k (non-blind deconvolution), i.e., for optimizing w.r.t. x_γ and ξ_γ. Specifically, ρ^* is an increasing function[3] and can be unbounded, in which case the MAP objective (27) always decreases with increasing ξ_γ, leading to the global minimum $\xi_\gamma \to \infty$ and $x_\gamma = 0$. Hence, the optimization will provide the trivial global minimum (a flat image) unless it is stuck at some local minimum. Even for bounded ρ^*, $\xi_\gamma(i)$ can be unbounded for $x_\gamma(i) = 0$ which will make the algorithm trapped at a local minimum. Thus, sparse MAP deconvolution based on iterative reweighting (jointly estimating x_γ and ξ_γ) can be unstable even in the non-blind case. On the other hand, in the VB approach, the term $\text{trace}\left(\text{diag}\left(\xi_\gamma\right) C_{x_\gamma}\right)$ implicitly enforces a regularization on ξ_γ: since C_{x_γ} is positive, increasing ξ_γ increases this term while $-\rho^*(\xi_\gamma(i)/2)$ decreases, such that the solution $\xi_\gamma \to \infty$ and $x_\gamma = 0$ is avoided.

However, these problems do not necessarily mean that all local optima in the MAP approach are useless; reasonable solutions can be obtained if some heuristic measures are taken. To address the second problem, the variables ξ_γ can be bounded from above by a positive number, which makes the MAP problem well-defined. Existing methods relying on sparse image priors, such as $\|x_\gamma\|_p^p$, $0 < p < 1$ in [18], or the semi-quadratic prior in [5], employ this bounding strategy although this problem is generally not recognized. In addition, these issues are the likely causes of the failure of the MAP estimate using the mixture-of-Gaussians prior reported in Fergus *et al.* [2], which noted that the MAP solution is often either the blurry image itself or a two-tone image. The image is estimated as the blurry input image when the blur is estimated as a delta function, which occurs because of the first problem. The two-tone image is most likely because most ξ_γ variables were driven to very large values (except possibly at the strong edges), resulting in extreme over-smoothing.

3.3 Bottom-Up Construction of Sparse Image Priors

So far, we started from an image prior definition using an analytical function ρ, and derived the estimation rules according to this function. This is the standard approach in most existing deconvolution methods, and is implicitly related to a "generative" view, i.e., the image prior reflects our prior belief in the natural image statistics. However, as mentioned above, for the prior to be super-Gaussian the only necessary condition is that $\rho'(s)/s$ is decreasing in $(0, \infty)$. Hence, instead of specifically assigning a function ρ and calculating $\rho'(\nu_\gamma)/\nu_\gamma$ to estimate ξ_γ, we can choose instead an arbitrary decreasing function κ for this estimation. In this way, we are not explicitly specifying an image prior, but achieving increased flexibility with a bottom-up construction while maintaining super-Gaussianity (hence sparsity) in the image prior. Such an approach leads to selective edge-preservation since small ξ_γ values correspond to small penalties on x_γ and will preserve them (and vice versa). For example, we consider a general form

[3] For $a \geq b \geq 0$, $\rho^*(a) = \inf_{x_\gamma(i)} \frac{1}{2} a\, x_\gamma^2(i) - \rho\left(x_\gamma(i)\right) \geq \inf_{x_\gamma(i)} \frac{1}{2} b\, x_\gamma^2(i) - \rho\left(x_\gamma(i)\right) = \rho^*(b)$. For some ρ (like log), ρ^* is strictly increasing with no finite asymptote.

$$\mathrm{E}\left[\xi_\gamma\right] = \kappa(\nu_\gamma) = \left(\frac{1}{g \otimes \nu_\gamma}\right)^p \tag{31}$$

where $p > 0$ and g is a filter. A variety of heuristics can easily be embedded through g to remove small gradients to combat noise and increase robustness (such as r-maps [7], bilateral filtering [13]), or to make strong gradients more pronounced (e.g., using shock filters [7], edge reweighting [5]). This construction can also be used to incorporate image properties not captured by a super-Gaussian prior, such as saturations. For such model discrepancies, cases where the decreasing property of κ is locally violated can also be considered. Finally, bottom-up construction opens the door for using spatially-varying priors (such as sparse priors around strong edges and less-sparse priors in texture areas).

Our formulation also provides a new interpretation of the mixture-of-Gaussians (MoG) prior used in Fergus *et al.* [2], where this prior is motivated as a sparse prior. In [2], four Gaussians are assigned to each pixel with corresponding ξ_γ^l, $l = 1, 2, 3, 4$, with prior values $\xi_\gamma^{l,0}$ estimated from training images. The VB inference scheme of [11] is employed to estimate ξ_γ^l as $\pi_\gamma^l \nu_\gamma^{-2} + (1 - \pi_\gamma^l)\, \xi_\gamma^{l,0}$, along with the corresponding mixture coefficients π_γ^l. In our modeling, this update rule can be obtained using $\rho(x_\gamma(i)) = \sum_l \pi_\gamma^l \log |x_\gamma(i)| + (1 - \pi_\gamma^l)\, \xi_\gamma^{l,0}\, x_\gamma^2(i)/2$. Thus, by jointly estimating ξ_γ^l and x_γ, Fergus *et al.* is in fact enforcing a mixture of super-Gaussians, which leads to sparse regularization on the image. The MoG prior was used in [2] only for blur estimation, but it can also be used for image estimation in our method via (24). Empirical results suggest, however, that a MoG prior is not necessary and using a single $\xi_\gamma(i)$ per pixel actually provides comparable or better results with simpler estimation rules (see Sec. 4).

4 Experiments

We implemented the proposed method using a pyramid coarse-to-fine approach where k and x_γ are estimated starting from downsampled images, and propagating the results up to the original resolution. We also adapt the noise variance σ^2 during the coarse-to-fine procedure by starting with a large value and gradually reducing it, as suggested in [9]. A `Matlab` implementation of the proposed algorithm is available online.

The proposed approach provides the flexibility of choosing the sparse image prior via the function ρ. To examine the effect of this choice, we used the synthetic dataset of [8] for a quantitative evaluation in terms of the sum-square-distances (SSD) between the deconvolved and original images. Figure 2 shows the cumulative SSD histograms for different selections for ρ (with $p = 0.8$ for $\|s\|_p^p$ and $\sigma_r = 0.9$ for the exp-prior), along with Levin *et al.*'s MAP$_k$ approach [9], which uses an MoG prior for kernel estimation and [18] for final image estimation. This method was shown in [9] to outperform the MAP based approaches of Shan *et. al.* [5] and Cho *et. al.* [6] in this dataset. We also show the ratios between the deconvolution error with the estimated kernel and the deconvolution error with the true kernel, which is used to normalize the influence of large kernels. It is evident that the log prior outperforms others both in blind and non-blind settings.

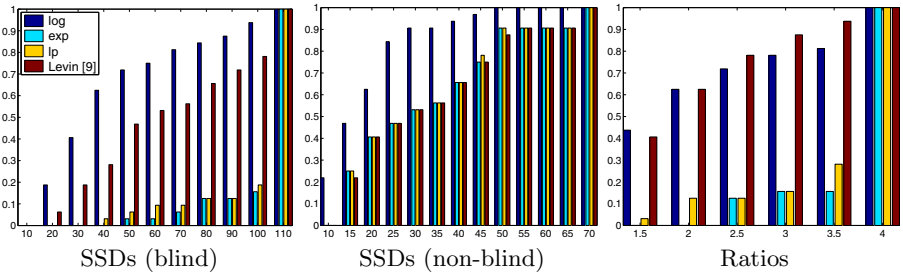

Fig. 2. Cumulative histograms of SSDs on the dataset of [8] for blind and non-blind deconvolution, and their ratios (right)

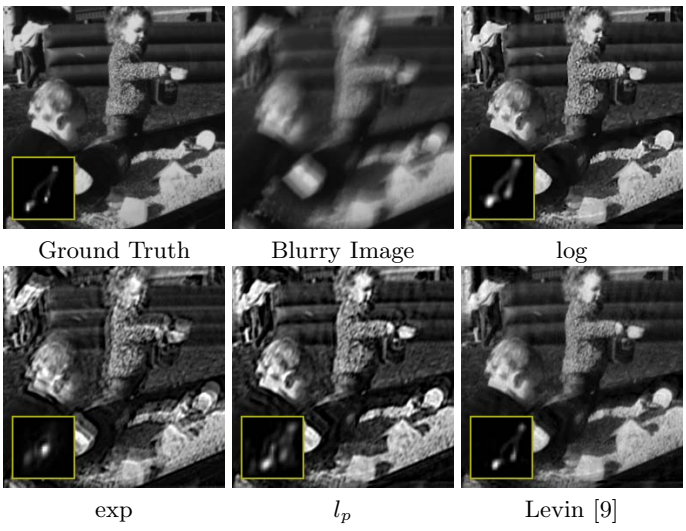

Fig. 3. Synthetic deconvolution results. Estimated kernels are shown in insets.

Also, its performance is in most cases better than that of [9]. Its success can be attributed to the enforced high sparsity, which improves the kernel estimation by distinguishing important edges while suppressing spurious features. Fig. 3 shows example visual results (full set of results can be found in the supplement), which shows that the low performance of the l_p and exp priors is mainly due to the poor kernel estimates.

Fig. 4 compares the proposed method with the log prior with previous methods on real images. While all methods perform well with subtle differences, the results of our method generally exhibit fewer artifacts and are more faithful to the original images. Our method is also considerably simple to implement and only requires tuning of one parameter (σ^2), whereas state-of-the-art deconvolution methods generally involve additional algorithmic steps (such as image filtering, kernel thresholding, etc.), which are not employed in our implementation but can easily be added for potential improvements in quality.

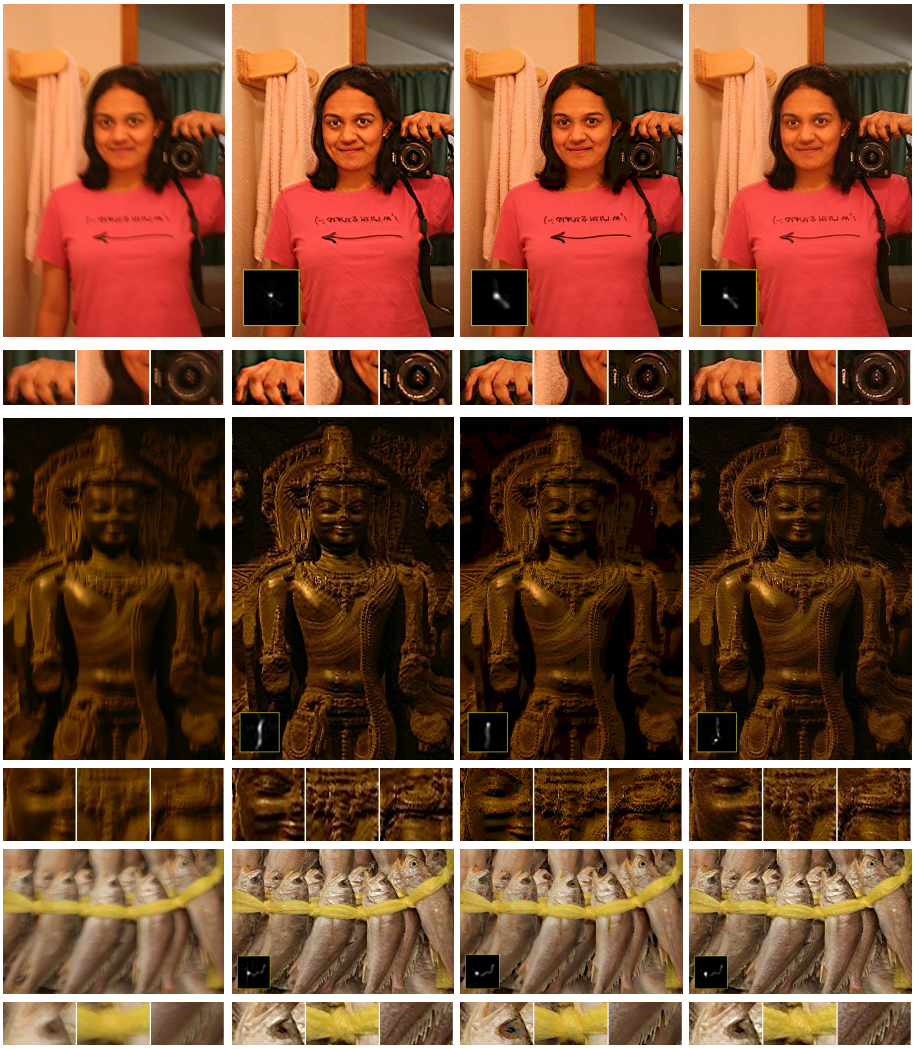

Fig. 4. Deconvolution results on real images. Estimated kernels are shown in insets. See the supplement for high resolution versions.

5 Discussion

This paper provides a systematic formulation of blind deconvolution using general sparse image priors. Any super-Gaussian prior can be used in this method with simple and efficient estimation rules. We also introduced a powerful new super-Gaussian prior. Our analysis of the proposed method showed that our method accurately addresses the major shortcomings of MAP (the delta kernel

and the flat image solutions) both for blind and non-blind deconvolution. We also showed that more flexible and powerful methods can easily be designed using our framework by modeling images locally with bottom-up construction of priors, which may prove useful in challenging deconvolution scenarios.

While a large family of image priors can be used within our method, some penalty functions such as scale invariant $\|\cdot\|_1/\|\cdot\|_2$ in [10] are not included. In addition, we only considered transforming the priors into Gaussian forms, but other options are also possible which may lead to faster methods. Finally, our method currently only addresses spatially-invariant blind deconvolution, which is limiting in practice. The formulations in [12,13] for representing spatially-varying kernels in terms of homographies can be employed to increase the applicability of the method. These are left as interesting directions for future work.

Acknowledgments. This work was partially supported by the Beckman Institute postdoctoral fellowship, "Ministerio de Ciencia e Innovacion" under contract TIN2010-15137, the Spanish research program Consolider Ingenio 2010: MIPRCV (CSD2007-00018), NSF Grant CCF-0916953, and a grant from the Department of Energy (DE-NA0000457).

References

1. Bishop, T.E., Babacan, S.D., Amizic, B., Chan, T., Molina, R., Katsaggelos, A.K.: Blind image deconvolution: problem formulation and existing approaches. In: Blind Image Deconvolution: Theory and Applications. CRC Press (2007)
2. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. ACM Trans. Graph. 25, 787–794 (2006)
3. Joshi, N., Zitnick, C., Szeliski, R., Kriegman, D.: Image deblurring and denoising using color priors. In: CVPR (2009)
4. Jia, J.: Single image motion deblurring using transparency. In: CVPR (2007)
5. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. ACM Trans. Graph. (SIGGRAPH) (2008)
6. Cho, S., Lee, S.: Fast motion deblurring. ACM Trans. Graph. (SIGGRAPH ASIA) 28 (2009)
7. Xu, L., Jia, J.: Two-Phase Kernel Estimation for Robust Motion Deblurring. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 157–170. Springer, Heidelberg (2010)
8. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Understanding and evaluating blind deconvolution algorithms. In: CVPR, pp. 1964–1971 (2009)
9. Levin, A., Weiss, Y., Durand, F., Freeman, W.T.: Efficient marginal likelihood optimization in blind deconvolution. In: CVPR, pp. 2657–2664 (2011)
10. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. In: CVPR (2011)
11. Miskin, J.W., MacKay, D.J.C.: Ensemble learning for blind image separation and deconvolution. In: Advances in Independent Component Analysis. Springer (2000)
12. Whyte, O., Sivic, J., Zisserman, A., Ponce, J.: Non-uniform deblurring for shaken images. In: ICCV (2010)
13. Hirsch, M., Schuler, C.J., Harmeling, S., Schölkopf, B.: Fast removal of non-uniform camera shake. In: ICCV (2011)

14. Palmer, J.A., Kreutz-Delgado, K., Makeig, S.: Strong Sub- and Super-Gaussianity. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 303–310. Springer, Heidelberg (2010)
15. Rockafellar, R.T.: Convex analysis. Princeton University Press (1996)
16. Black, M.J., Rangarajan, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. IJCV 19, 57–91 (1996)
17. Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological) 36, 99–102 (1974)
18. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. ACM Trans. Graph. 26 (2007)
19. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
20. Likas, A.C., Galatsanos, N.P.: A variational approach for Bayesian blind image deconvolution. IEEE Trans. on Signal Proc. 52, 2222–2233 (2004)
21. Babacan, S.D., Molina, R., Katsaggelos, A.K.: Variational Bayesian super resolution. IEEE Trans. Image Proc. 20, 984–999 (2011)

Bayesian Blind Deconvolution with General Sparse Image Priors Supplementary Technical Note

S. Derin Babacan[1], Rafael Molina[2], Minh N. Do[1], and Aggelos K. Katsaggelos[3]

[1]University of Illinois at Urbana-Champaign, [2]Universidad de Granada, [3]Northwestern University

In the following, the equation and table numbers are denoted with preceding "A-". The ones without the preceding "A-" refer to the main paper.

Concave Conjugate Formulation

Given the function $\rho(\sqrt{x_\gamma(i)})$ which is concave and increasing on $(0,\infty)$, we have the concave conjugate pair

$$\rho\left(x_\gamma(i)\right) = \inf_{\xi_\gamma(i)>0} \frac{1}{2}\,\xi_\gamma(i)\,x_\gamma^2(i) \;- \rho^*\left(\frac{\xi_\gamma(i)}{2}\right) \tag{A-1}$$

$$\rho^*\left(\frac{\xi_\gamma(i)}{2}\right) = \inf_{x_\gamma(i)} \frac{1}{2}\,\xi_\gamma(i)\,x_\gamma^2(i) - \rho\left(x_\gamma(i)\right)\,. \tag{A-2}$$

Taking the derivative of the right hand side of (A-1) w.r.t. $\xi_\gamma(i)$ and setting it equal to zero, we obtain

$$x_\gamma^2(i) = \rho^{*'}\left(\frac{\xi_\gamma(i)}{2}\right)\,. \tag{A-3}$$

Next, taking the derivative of the right hand side of (A-2) w.r.t. $x_\gamma(i)$ and setting it equal to zero, we have

$$\frac{\rho'\left(x_\gamma(i)\right)}{x_\gamma(i)} = \xi_\gamma(i)\,. \tag{A-4}$$

The equalities in (A-3) and (A-4) specify the optimal value of $\xi_\gamma(i)$ for a specific value of $x_\gamma(i)$.

This representation is used in the paper in order to derive the cost-function for VB. Since the distribution $p(\xi_\gamma)$ is degenerate, using (A-3) in (A-4), the VB estimate of $\mathrm{E}\left[\xi_\gamma(i)\right] = \frac{\rho'(\nu_\gamma(i))}{\nu_\gamma(i)}$ in (23) can be written as the solution of

$$\mathrm{E}\left[\xi_\gamma(i)\right] = \arg\min_{\xi_\gamma(i)} \frac{1}{2}\,\xi_\gamma(i)\,\nu_\gamma^2(i) \;- \rho^*\left(\frac{\xi_\gamma(i)}{2}\right) \tag{A-5}$$

which is used in deriving the cost-function for the VB approach (26) in Section 3.2 of the main paper.